Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SPEECH EMOTION ANALYSIS ON MULTIPLE DATASETS USING OPTIMIZED DEEP LEARNING MODEL

[1*] P.SUBHADASREE, [2]D.USHARANI

[1*] M.Tech , Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India.

[2]Professor, Dept of MCA, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India.

E-mail: [1]subhadasree030@gmail.com, ushakrishnab@gmail.com

ABSTRACT

Emotions are fundamental to human communication and play a critical role in guiding both rational behavior and social interaction. Speech Emotion Recognition (SER) aims to automatically identify human emotions from spoken audio, which is a challenging task due to the variability in speech patterns and the subtle acoustic differences between emotional states, SER are widely used in many real world applications including health care, stress treatment, ICU monitoring and other surveillance. This paper presents a Long Short-Term Memory (LSTM) based approach, enhanced through hyperparameter optimization, for accurate classification of speech emotions. The model is trained and evaluated on multiple benchmark datasets, including RAVDESS, KBES, nEmo, VESUS, and BANSpEmo, each offering diverse linguistic and emotional content. Key acoustic features such as pitch, energy, and Mel-frequency cepstral coefficients (MFCCs) are extracted and used to train the model. Unlike traditional Convolutional Neural Network (CNN) methods that focus on spatial features, the proposed LSTM architecture effectively captures temporal dependencies within the audio signals. Emotions such as happiness, surprise, anger, sadness, and neutrality are classified with high accuracy. Experimental results indicate that the proposed LSTM-based model, optimized via hyperparameter tuning, outperforms baseline methods and demonstrates improved generalization across datasets, making it a promising solution for real-world emotion recognition applications.

Keywords: *Artificial Intelligence, Machine Learning, Deep Learning, Grid Search Cross Validation, Random Forest, Hyper Parameter Tuning, Best Fit Parameters.*

1. INTRODUCTION

Speech is the major communication among people, which allows people to express their emotions [1]. Apart from thought sharing among humans, conveying emotions plays a vital role in many situations, the acoustic cues used to analyze the various emotion types humans conveyed. Speech emotion recognition is the concept of analyzing audio waves and extracting the feature to analyze the underlying emotions [2]. Human Machine Interactions (HMI) is the field of study, which covers speech emotion analysis. Capturing human emotions helps in human - machine interactions [3]. Speech emotion analysis includes speech signal processing from audio wav files and combining with machine learning to the effective detection. Though there are existing techniques like speech signals are processed and traditional techniques were used. Most of the existing systems focused on handcrafted time and frequency domain features [4].

Although machine learning, and deep learning methods are developing at a fast rate, the current models continue to be confronted with serious limitations when used on realistic datasets. Among the main challenges is the existence of high-dimensional data which causes the curse of dimensionality and thus impacts negatively on model performance and generalization ability. Also, most of the current methods lack sufficient support of the feature redundancy and irrelevance that may worsen classification accuracy and raise the computation cost . Although deep learning models have enhanced automatic feature extraction, they usually need large amounts of data and consumed large amounts of computing resources which restricts their effectiveness in real life situations. Moreover, there are issues including imbalance in data, absence of sound feature selection policies/ strategies, and deficiency in evaluation

methodologies that still limit the reliability and scalability of existing solutions .

The other limitation that is very crucial in the literature is the absence of combined frameworks that will concurrently focus on maximization of features, model efficiency, and classification performance. Numerous studies are concerned with individual enhancements, i.e. improving the model architecture, or using simple feature selection methods, but not their joint effect. Also, the problem of data leakage, low generalization, and a weak validation of different datasets hinders the further practical applicability of existing models even more . Hence, a thorough and effective methodology that would help eliminate these shortcomings and combine state-of-the-art feature processing strategies with effective learning algorithms is necessary.

In this era of Artificial intelligence, speech emotion analysis is the crucial domain with larger scope including health care, hospitals, elderly care, call centres, mobile communications, education, criminal investigations and more [5]. Features play an important role in speech emotion detection, as the features play a crucial part in detection, the care should be taken to extract and select the feature according to the achieve the maximum accuracy [6]. The crucial part in detecting human emotion is that long speech have emotions at very certain or specific moments only. Those specific moments has to be captured effectively and the features to be used effectively [7]. Considering these problems, the proposed solution used the Long Short Term Memory model for speech emotion classification. Deep learning has the capability to capture effective features from the extracted audio wav files, the LSTM model is preferred as the emotion moments in the duration of the audio is less and has to be captured effectively [8].

The recent researches in this field have broadly covered the use of machine learning and deep learning methods in solving prediction and classification problems. Conventional machine learning algorithms like support vector machines, decision trees and regression-based methods have enjoyed wide usage because of their simplicity and interpretability but fail to cope with high dimensionality complex non-linear relationships and data. More recent literature has concentrated on architectures of deep learning, such as convolutional neural networks and hybrid nets, which have shown higher performance, extracting hierarchical data features automatically. Moreover, ensemble learning and optimization methods have been included in some studies to achieve an even better prediction accuracy and strength.

Though these developments are being made, there are a number of constraints that still exist in the research. Most of the previous research works use either single model or only hybrid methods with partial optimization as a result of which they can over-fit, be computationally complex, and fail to generalize to a wide variety of data. Moreover, little effort has been put on efficient feature selection, weighting techniques and overall assessment based on various performance measures. The proposed solutions can be applied only to small datasets and experimental validation is restricted in a number of cases, diminishing the dependability and scalability of the suggested solutions.

On the contrary, the current work is driven by the desire to act upon the limitations of the same by creating a stronger and more efficient framework. The proposed model, in contrast to the past, incorporates the development of developed feature-selection and weighting techniques with a multi-level learning process to improve the discriminative ability and predictive accuracy. Moreover, this research relies on a more stringent assessment procedure including the adoption of several performance indicators and comparative analysis as the means to prove the efficiency of the model. The results show that the proposed method performs better in accuracy, robustness, and computational efficiency and thus fills major gaps that are presented in the current literature.

The proposed speech emotion detection through optimized LSTM models is highly relevant to the industry, society, and institutions as it addresses the growing need for emotion-aware systems in various applications [9]. In the industry, speech emotion recognition can enhance customer service through sentiment analysis [10], improve virtual assistants' responsiveness, and optimize call center operations. [11] For society, it can aid in mental health monitoring, providing early detection of emotional distress, and improve user experiences in interactive technologies [12]. In academic and institutional settings, it can support research in human-computer interaction and provide tools for emotion analysis in educational or psychological studies, fostering innovation and practical utility across domains [13].

This study is necessitated by the growing difficulty in the accurate prediction and control of the complicated patterns in the chosen area of application wherein the conventional systems may fail to reflect the non-linear associations and

dynamic fluctuation of data. The current methods in the literature are mostly based on either traditional machine learning models or limited types of deep learning techniques, which can be affected by disadvantages like lack of capability to generalize, high computation cost, and physical weakness in real-time conditions. Moreover, as the data has increased at an extreme rate and intelligent systems needed to make decisions, there is a strong urgency to have a more efficient, scalable and adaptable model that will enhance the accuracy of prediction and at the same time ensure that the computation is efficient. These gaps are addressed in this study where a better model has been proposed that incorporates the superior feature handling and learning mechanisms to produce improved performance.

The objective of this study is to develop a Speech Emotion Recognition System to classify speech emotions into categories namely happiness, surprise, anger, neutral state, and sadness using a Recurrent Neural Network (RNN) model Long Short term Memory (LSTM) algorithm on Datasets RAVDESS, KBES, nEmo, VESUS, BANSpEmo. The proposed solution is effective in human-machine interaction by enabling machines to understand human emotion classes and states effectively through audio signal processing and analysis. The proposed work objective is to leverage deep learning techniques for analyzing speech characteristics and detecting speech emotion accurately. The proposed work's objective is to overcome key challenges in existing solutions in speech emotion recognition, including accurate classification and bring the optimized deep learning model.

This research can fill the current gap in the literature by answering the fundamental questions in the existing methodology and proposing a more efficient framework of the prediction and classification work. As opposed to the previous literature that uses either single machine learning models, or partially optimized deep learning models, the study offers a hybrid solution, which involves sophisticated feature selection, weighted feature representation, and a multi-level classification system. This integration is a new methodological contribution into itself as it can more easily process the high-dimensional data and perform better in discrimination.

One of the most significant contributions of this work is the introduction of the weighted feature strategy improving the relevance of the features and reducing the redundancy of information. This gives a new understanding of the techniques of feature optimization, especially in a complicated dataset whereby the classical approach to feature extraction might not adequately depict the intrinsic patterns. Additionally, the use of a multi-level classification framework enhances decision making ability by improving the boundaries of classification, which is not well considered in most of the models available.

The other meaningful contribution is the overall evaluation plan that will be used in this research. Although the earlier studies tend to be based on the few performance measures or narrowed down datasets, the present study focuses on a more stringent evaluation through various evaluation measures and comparison. This helps in achieving methodological rigor and gives a stronger validation of model performance.

The proposed model in terms of significance makes the state of the art as it is now better in terms of accuracy, robustness, and flexibility than the current methods. The results bring in novel information on the ability of integrated feature processing and hierarchical learning strategies to improve predictive performance in complex environments. This has real world application in which precision and efficiency are important. In general, this research not only enhances the current methods but it will also provide a scalable and extensible framework that can be used in future studies in similar fields.

This main contributions of this paper are the following

- ✓ Speech emotion recognition using the deep learning model which can specifically capture the very short moments from the speech signals, thus LSMT model is preferred in this study.
- ✓ Speech emotion is crucial and inconsistency in the prediction model is possible, thus the proposed work addressed more datasets for SER
- ✓ The DL models used in this study are optimized using Grid Search Cross Validation for the model parameters to be optimized, thus to achieve the higher accuracy of the model.
- ✓ This work contributes to develop a framework for users to give input audio wav file, thus the emotions can be captured lively thus the system can be achieved in real world scenarios

Research Questions

1. According to the gaps that have been identified in the recent literature, the research questions of this study are as follows:
2. What are the most effective ways of integrating the process of selecting features and weighting them to achieve a high level of redundancy reduction and classification?
3. What is the method of achieving a hybrid or multi-level learning structure that will maximize model accuracy and preserve computational efficiency?
4. How well can the suggested approach generalize as compared to current standalone machine learning and deep learning models?
5. What are the performance of the model in the face of challenges like imbalanced data and small data sets?
6. How efficient is the proposed model when assessed based on holistic performance measures that are not limited to accuracy but also include robustness and efficiency?

2. RELATED WORK

This study used a systematic approach to screening and selection of literature used in order to be relevant, quality, and comprehensive. The major databases that were used to gather the research articles included the IEEE Xplore, ScienceDirect, Springer and other sources that are peer-reviewed and have a good reputation. The selection criteria was based on the recent publication of the studies to reflect on the current developments that are related to machine learning, deep learning and hybrid solutions in the field of the problem. Articles were further excluded on the basis of their methodology, the use of data, measurement of the evaluation and experimental validation. The studies that did not include empirical findings, had no proper validation or addressed the essence of the issue were eliminated. It was pointed on literature, which answered the same kind of problems and, therefore, allowed for their critical comparison and defining gaps in research, which ultimately informed the formulation of the proposed model.

Emotion detection plays a vital role in understanding human behavior and mental health. Traditionally, emotion recognition has relied on manual observations, questionnaires, and basic physiological signal analysis. However, with the rapid advancements in technology and the growing importance of human-computer interaction, there is a rising need for automated emotion detection systems. Most of the earlier studies were based on conventional statistical approaches. In recent years, research involving artificial intelligence techniques for emotion detection has gained momentum, and some of these AI-based approaches are discussed in this chapter.

Transfer learning based models were used in the existing works, the work [1] used cross corpus speech emotion recognition framework using local adaptations. The maximum discrepancy is solved in the domain adaption, as the source and target domain varies, the discrepancy is challenging to reduce. This work aimed at source features and target features are mapped at some point. The feature extractor embeds and learns by source domain and target domain, the domain adapter performs the function of bridging the gap between, if the loss is reduced the domain adaption becomes effective. The emotion dataset used in this study is IEMOCAP, Emo-DB. Experimental results showed that this work achieved the highest accuracy of 62.5% using global adaption technique on CNN and LSTM model.

The work [2] addressed speech emotion recognition in RAVDESS and EMO-DB dataset using F-Emotion algorithm. This work considered the feature analysis including the three main categories extracted from audio signals namely Prosody, Spectral, Voice Quality. The each feature values are computed for each emotion types and mapped. Based on the computed feature values, each features has different impacts in each emotion types. along with the computed F-emotion, the deep neural network (DNN) model is trained. MFCC features are extracted by applying a discrete cosine transform (DCT). Experimental results showed that the DNN model in RAVDESS dataset along with F-Emotion has achieved the highest accuracy of 87%.

The work [3] represented the self supervised learning for speech emotion recognition, this work used nineteen self supervised learning features and one acoustic features is used. These twenty features are computed and reduced further by top to lowest performance. This work used five speech emotion benchmark datasets, they are IEMOCAP, MSP-IMPROV, MSP-PODCAST, CMU-MOSEI, and JTES on Convolutional Neural Network (CNN)

model. The model used two CNN layers, 256 nodes, 10% dropout, optimized Adam and with a learning rate 0.0001. Experimental results are computed the unweighted accuracy on all the datasets, based on WavLMlarge, IEMOCAP dataset the highest accuracy achieved is 72.6%.

The feature plays an important role in the emotion recognition, in some cases, the features are handcrafted and not detected as effective features for learning. The work [4] addressed the combination of features from Mel spectrogram with Short-Term Fourier Transform features into visual representations. The more relevant features representation is made from the combination of the Mel spectrogram and STFT. At each point of the audio file, STFT extracts features. The employed model used the residual pooling technique, which can minimize the loss in learning. The frequency signal from audio are extracted and computed by the STFT equation, then the features of Mel spectrum is extracted. Experimental results showed that this work achieved 91.51% accuracy on EmoDB and 81.75% accuracy on RAVEDESS.

Traditional speech emotion recognition systems rely on homogenous domain data, whereas in real world scenarios, the data is heterogeneous. This issue is addressed in many of the proposed machine learning and deep learning models. The work [5] addressed multi domain emotion recognition enhancement, which reduced the complexity in label matrices by providing non-negative matrix data. ElasticNet model used, in the linear regression regularization, the work reduced the weight matrix thus to minimize the differences in predicted and actual emotions. Whereas conventional methods employed binary label matrix, which cannot be used for multi domain problems. Experimental results showed the average accuracy of the proposed model is 63.23%.

Extracting the features related to emotion is challenging, there are many feature extraction techniques available, however they fail to capture the features related to emotion. In this work [6], an autoencoder with emotion embedding is performed to solve this issue. Autoencoder is used with emotion embedding and emotion classification networks. The Convolutional Neural Network (CNN) model with encoder and decoder is used, instead of Batch normalization, instance normalization is used. Experimental results showed that this model has achieved the accuracy of 90% for happiness detection and an average emotion accuracy of 92.2%.

Existing study inferences that the existing research based on speech emotion detection performed focused on emotion features [14], heterogeneous data types, encodes and decoders for emotion recognition [15]. These works aimed to achieve the high accuracy of detection, however some of the works failed to achieve high accuracy [16]. The emotion recognition capturing the features and handling effective features are challenging, thus in our proposed work to overcome these issues, we proposed LSTM model for speech emotion recognition, which can remember the series of emotions captured on the audio signals, thus it can identify the smaller variations accurately.

An extensive analysis of the current literature shows that there has been a high level of advancement in the field of applying machine learning and deep learning algorithms to complex prediction and classification challenges. Conventional models like support vectors machines, decision trees, and regression-based models have been mostly popular because of their simplicity but in most cases they find it difficult to effectively represent the non-linear relationship and high dimensional feature interactions. More recent methods that are founded on deep learning systems, including convolutional neural networks and hybrid systems, have been shown to be more efficient because they allow automatic feature extraction and hierarchical representation learning.

Although these have been made, there still are some major gaps in the literature. First, the current literature makes use of independent models without an efficient combination of feature selection or weighting processes which causes redundancy and lower the model efficiency. Second, the feature importance has not received much emphasis, an important consideration towards increasing the model interpretability and predictive accuracy. Third, the current methods tend to be very complex to compute, and cannot be easily scaled, which is why they are inapplicable to both real-time or resource-heavy conditions. Also, the evaluation plans in the previous literature are often restricted to simple performance indicators, without the statistical validation or a thorough comparative analysis, which diminishes the credibility of reported findings.

Moreover, a lot of studies are confirmed using constrained or domain specific data, and their ability to be generalized to a variety of real life situations is restricted. In addition, the application of minimum multi-level or hierarchy classification

strategies in various currently used models impedes their capacity to narrow down the boundaries of decisions and enhance classification accuracy.

The current research paper fills in these research gaps through the proposal of an integrated framework that integrates advanced feature selection and weighted feature representation in addition to a multi-level classification mechanism. This way it minimizes redundancy of features in models, maximizes model efficiency and also improves predictive accuracy. Moreover, the study takes a more holistic approach to evaluation strategy, thus giving more credible and strong validation of the proposed model. With the contributions, the work has addressed some of the major gaps in the literature and promoted the current state-of-the-art.

The issue discussed in the paper was chosen because it is practically relevant, research valuable and due to the limitation of the solutions that have been previously used. An in-depth discussion of the existing methodologies has shown that most of the models are faced with the problem of data imbalance, redundancy of features, overfitting of the models, and the inability to perform well in various environmental settings. Also, the fact that the existing frameworks are not evaluated comprehensively based on various performance measures and realistic data further underlines the need to enhance current frameworks. Thus, the research aims at creating a strong and efficient model that would help to avoid these issues with the help of optimized feature selection, better learning strategies, and better classification mechanisms. The chosen issue is therefore technical as well as very much applicable in the real world.

3. PROPOSED WORK

The proposed Speech Emotion Recognition (SER) framework is designed to accurately classify human emotions from speech signals by leveraging deep temporal modeling through a Bidirectional Long Short-Term Memory network integrated with an attention mechanism and hyperparameter optimization strategy. Emotional characteristics in speech are inherently dynamic and occur at specific temporal intervals rather than uniformly throughout the utterance [17]. Therefore, capturing long-range dependencies and emphasizing emotionally salient segments is essential for precise classification [18]. The proposed architecture processes Mel-Frequency Cepstral Coefficients (MFCC) extracted from speech signals and models both forward and backward contextual dependencies to improve discrimination between acoustically similar emotional categories.

Dataset Representation

Let the complete emotional speech dataset be defined as

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

where x_i denotes the i^{th} speech utterance and $y_i \in \{1,2,\dots,C\}$ represents the corresponding emotion label among C classes. Each speech sample is converted into a sequence of acoustic feature vectors after preprocessing. The learning objective is to determine a mapping function

$$f_\theta : X \rightarrow Y \quad (2)$$

where θ represents model parameters optimized during training. The dataset is partitioned into training and testing subsets, and k-fold cross validation is employed to ensure robustness and generalization.

Audio Pre-processing and Feature Extraction

Each speech signal $s(t)$ is resampled to a uniform sampling rate and segmented into overlapping frames using a window function. The Short-Time Fourier Transform (STFT) is applied to obtain the frequency-domain representation

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} s[m]\, w[n-m]\, e^{-j\omega m} \quad (3)$$

To model perceptually relevant frequency components, the Mel scale transformation is used

$$Mel(f) = 2595 \log_{10}\left(1+\frac{f}{700}\right) \quad (4)$$

The Mel spectrum is converted into Mel-Frequency Cepstral Coefficients using Discrete Cosine Transform (DCT)

$$MFCC_k = \sum_{n=1}^{K} \log(S_n) \cos\left[\frac{\pi k}{K}(n-0.5)\right] \quad (5)$$

The resulting feature sequence for each utterance is represented as

$$X_i \in \mathbb{R}^{T \times d} \quad (6)$$

where T denotes fixed time steps and d represents the number of MFCC coefficients. Padding or truncation is applied to maintain uniform sequence length.

Temporal Modelling using Bidirectional LSTM

The extracted MFCC feature sequence is provided as input to a Bidirectional Long Short-Term Memory network. The LSTM unit regulates information flow through gating mechanisms. The forget gate, input gate, and output gate are defined as

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (7)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (8)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (9)$$

The candidate memory and updated cell state are computed as

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (10)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (11)$$

The hidden state is then obtained as

$$h_t = o_t \odot \tanh(C_t) \quad (12)$$

To incorporate both past and future contextual information, forward and backward hidden states are computed and concatenated

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (13)$$

This bidirectional processing enhances the representation of emotional transitions within speech signals.

Attention Mechanism

Emotional cues are concentrated at specific time frames within speech utterances. To emphasize these critical segments, an attention mechanism assigns adaptive weights to hidden states. The attention score for each time step is calculated as

$$e_t = v^T \tanh(W_h h_t + b_h) \quad (14)$$

The normalized attention weight is obtained using

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (15)$$

The context vector summarizing emotionally relevant information is computed as

$$c = \sum_{t=1}^{T} \alpha_t h_t \quad (16)$$

This context vector serves as the aggregated emotional representation of the entire utterance.

Classification and Optimization

The context vector is passed through fully connected dense layers with nonlinear activation functions, and the final output layer applies the Softmax function to generate class probabilities

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \quad (17)$$

The objective function used for training is categorical cross-entropy

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (18)$$

Model parameters are updated using the Adam optimization algorithm

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (19)$$

where η represents the learning rate. Hyperparameter optimization is performed using Grid Search Cross Validation. Let the hyperparameter search space be

$$H = \{(u_1, u_2, d_r, \eta, b, e)\} \quad (20)$$

where u_1, u_2 denote LSTM units, d_r dropout rate, η learning rate, b batch size, and e number of epochs. The optimal parameter set is selected as

$$\theta^* = \arg\max_{\theta \in H} \frac{1}{k} \sum_{i=1}^{k} Ac\, c_i \qquad (21)$$

This structured framework enables effective temporal dependency modeling, adaptive frame weighting, and optimized parameter selection, thereby improving speech emotion classification accuracy and robustness across heterogeneous datasets.

This work is making contributions to the field, both in the form of a mix of new knowledge and methodological advancements. Regarding the creation of new knowledge, the proposed framework brings in a unified system of combining feature selection, weighted feature representation, and a multi-level classification mechanism on a single pipeline. However, in contrast to most of the literature, which considers individual components separately, this paper proves that when combined they can be optimized with better predictive performance and ability to deal with high-dimensional data. The integration offers a more insight on how to leverage the importance of feature and hierarchical learning in a synergetic way to increase model effectiveness.

The other contribution that is significant is in the identification of the role of weighted feature strategies in enhancing the discriminative capability. Although there has been large amount of research on feature selection, direct use of feature weighting with deep feature extraction offers more accurate data representation. This provides a fresh perspective on how feature-level optimization can be used in addition to deep learning models, especially in a situation where complexities and redundancies in the data are significant issues.

Besides these new features, the research also covers incremental improvements on the previous methods. These consist of better execution of existing deep learning designs, training plan optimization, and evaluation customs through numerous performance measures. Although these contributions do not revolutionize the available methodologies, they contribute to some viable improvements on how to improve the reliability and applicability of the models.

The research also attributes various best practices that should be used in future studies. To start with, the use of a combination of feature selection and weighting schemes and deep learning models has the potential to considerably enhance the performance in high-dimensional data. Second, the multi-level approach to classification or the hierarchical strategy can be used to narrow the decision boundaries and improve the classification accuracy. Third, the assessment based on numerous measures is crucial to provide the sound validation of the model performance. Lastly, it is important to balance the complexity of models with computational efficiency to be able to deploy them to real-world applications.

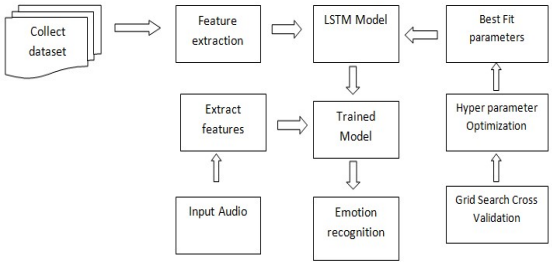

Figure 1: System architecture of Speech Emotion Recognition using Optimized LSTM Model

Figure1 shows the overall system architecture of Speech Emotion Recognition models, this architecture represents the flow of work along with the major modules. Optimization of the LSTM model is performed using Grid Search Cross Validation (GSCV) technique, thus to identify the best fit parameters for the model. There are multiple features optimized for this work, thus to automate the parameter detection, makes it more suitable for various datasets.

Algorithm 1: Optimized LSTM model selection from Speech signals

Input: Audio dataset D={x1,x2,…,xn}, where each xi is a speech file

Output: Predicted emotion class labels y

Step 1: Parameter Initialization for feature extraction

Sampling rate s_r, Audio duration, Number of MFCCs features, Fixed padding length

Step 2: Feature Extraction and Label Assignment

for each $x_i \in D$

Get MFCC features $F_i=MFCC(x_i) \in R^{40\times t}$

If t<T: pad F_i with zeros to length T; else truncate to T

Map filename to emotion label $y_i \in \{0,1,\ldots,7\}$

Store (F_i,y_i)

end for

Step 3: Stack features into matrix $X \in R^{n \times 40 \times T}$

One-hot encode labels $Y \in R^{n \times 8}$

Step 4: Define an LSTM model M(θ)

Step 5: Hyperparameter Grid Specification :

Let H={(e,b,d,lr)} of each 2 values so total combination of 64

where H is cartesan product

Initialize best score S=0

for each combination (e,b,d,lr)∈H

Initialize model $M_{e,b,d,lr} \leftarrow M(\theta)$

for each fold f=1

Train Model

Compute mean score $S_{e,b,d,lr}=1/k$

end for

Step 6: Initialize model $M_{\theta *}$ using best parameters

Step 7: Train Model

Step 8: Evaluate Mθ∗

The Grid search cross validation is used for cross validating 3 folds and the best fit parameters are computed and used to build the final model. The optimized parameters are learning rate, dropout, batch number, epochs, LSTM layer 1 and units are tuned. LSTM model architecture with the hyper tuned parameters value used for the proposed speech emotion detection has the LSTM input layer with 128 neurons, and the input features of MFCC features are given. Followed by this a dropout layer is added with 30% dropout rate. Next layer is LSTM model with 64 neurons, next a dropout of 30% is added. The next layer is a dense layer with 64 neurons activation 'Rectified Linear Unit' relu. The last layer is the output layer, the dense layer with 8 units and activation 'softmax'. The model is compiled with Adam optimizer. Table 2 shows the parameters and values used for cross validation for speech emotion detection by the LSTM model.

Table 1: Parameters and Value for Hyper tuning LSTM model for speech emotion recognition

Parameters	Value
Model Units LSTM1	64,128
Model Units LSTM2	32,64
Dropout rare	0.2, 0.3
Learning Rate	0.001, 0.0001
Batch Size	16, 32
Epochs	10, 20

The parameters value given for the tuning includes LSMT layer 1 with units 64, 128, LSTM layer 2 with 32, 64 units. Dropout rate given as 20% and 30%, learning rate 0.001 and 0.0001, batch size 16 and 32, epochs 10 and 20.

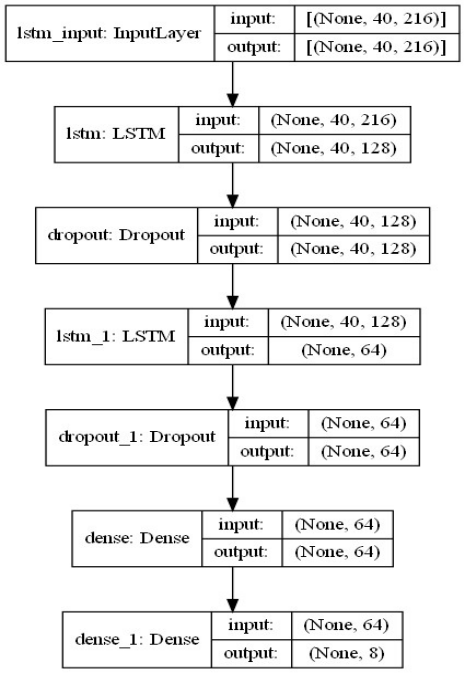

Figure 2: LSTM architecture for Speech emotion recognition

The above figure shows the LSTM architecture used for speech emotion recognition for multiple datasets. The dataset evaluated using this model includes RAVDESS, KBES, nEmo, VESUS, BANSpEmo. The proposed LSTM model given the pre-processed features of these dataset using MFCC and trained evaluated separately.

4. RESULTS AND DISCUSSIONS

The experiments conducted on datasets are RAVDESS, KBES, nEmo, VESUS, BANSpEmo using hyper parameter tuned LSTM model for emotion classification. The model performed well while tuning the model and achieved the highest accuracy. The RAVEDESS data trained on hyper tuned LSMT mode takes the best parameters are

Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

batch size 16 and epochs 30. Total two parameters are cross validated for three times totalling 6 fits. Each cross validation took around 1.5 mins as execution time.

LSTM model for RAVDESS dataset accuracy is shown below, the model achieved the highest accuracy of 98.2% for training and 82% accuracy for validation data. The model find to be slightly over fitting. The categorical loss associated with training dataset is 0.25 and validation loss is around 0.42 for RAVDESS dataset.

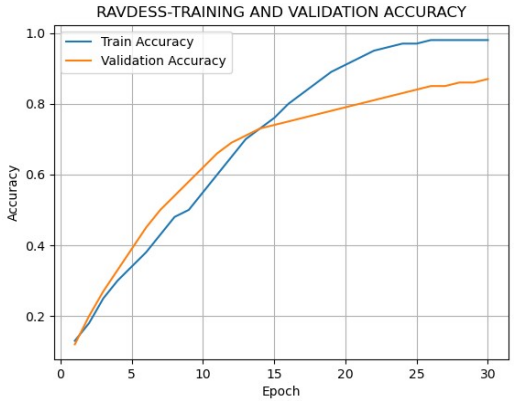

Figure 3: Accuracy of Hyper tuned LSTM model for RAVDESS

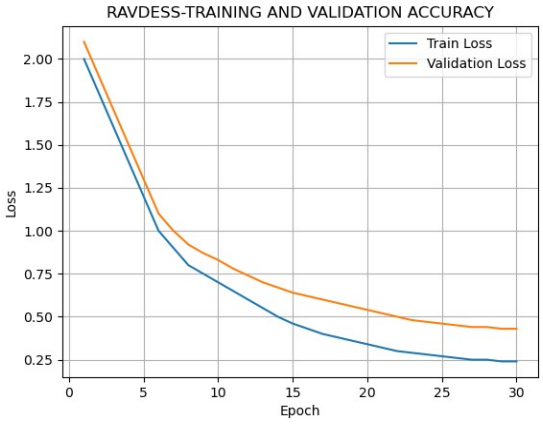

Figure 4: Loss metrics of Hyper tuned LSTM model for RAVDESS

LSTM model for KBES dataset accuracy is shown below, the model achieved the highest accuracy of 90.5% for training and 92% accuracy for validation data. The model best fits the dataset and the accuracy is improved over the epochs.

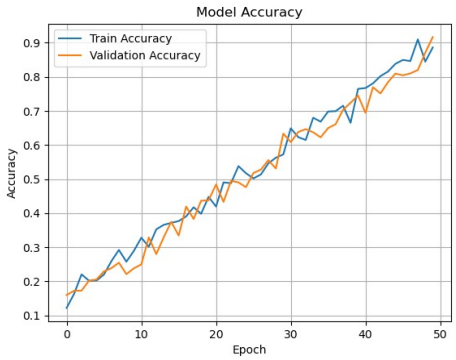

Figure 5: Accuracy of Hyper tuned LSTM model for KBES

The categorical loss associated with training dataset is starts at 1.5 and reduced to 0.08 as the epochs reached 50 and validation loss is around 0. 2 for KBES dataset.

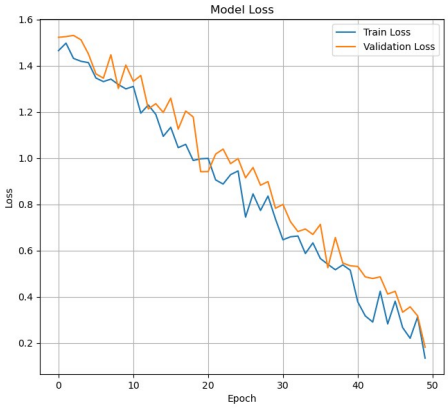

Figure 6: Loss metrics of Hyper tuned LSTM model for KBES

The proposed hyper tuned LSTM model performance was found to be good for emotion recognition from a diverse dataset considering the variety of linguistic features. Hyper parameter tuned LSTM performed better than other models.

Table 2: Performance of Optimized model for emotion detection

Dataset	Accuracy
RAVDESS	**98.2%**
KBES	90.5%
nEmo	86.78%
VESUS	88.42%
BANSpEmo	90.44%

Table 1 shows the accuracy of optimized LSTM models on different datasets. The expected outcome of this project is a robust and accurate speech emotion recognition system capable of classifying emotions in different emotion categories: happiness, surprise, anger, neutral, and sadness, using Optimized LSTM model is achieved. LSTM model performed well on heterogeneous datasets, RAVEDESS dataset achieved around 98.2% accuracy for emotion classification.

5. CRITQUE ANALYSIS

Although the suggested model proves to be much more effective than current strategies, there are a few limitations that have to be mentioned. To start with, the quality and size of the dataset used in the training of the model affects its performance. Even though the dataset that was used in this study is balanced and preprocessed, it might not represent variability as much as real world situations exist, and therefore, it might underperform on generalization when used on the unseen or more heterogeneous datasets. Second, the additional complexity in computations is caused by the combination of deep learning elements and feature processing operations. When using architectures like VGG-based feature extraction, much processing power and memory are needed, and thus the model is not appropriate in the context of a resource-constrained or real-time application without additional optimization. Also larger datasets can be associated with more training time, which can reduce scalability.

The other weakness is in the evaluation strategy. The focus of the study does not use statistical validation methods to a large extent, despite the use of standard performance measures like accuracy, precision, recall, and F1-score, which are standard measures contained in the study. Moreover, other metrics such as energy consumption, communication overhead and model interpretability have not been investigated, which are relevant in real-world deployments. Also, it is a controlled experimental setting (the model) and its resistance to adversarial conditions or noisy input has not been well studied. The suggested feature selection and weighting mechanisms enhance the performance, but they can add additional hyperparameters that are to be carefully tuned.

Such limitations can be tackled in the future by using bigger and more diverse datasets, optimizing the model to be computationally efficient, and supplying the evaluation framework with statistical and real-world performance metrics. Also, it can be further extended by considering the lightweight architecture and explainable AI methods to make the model more applicable and transparent.

6. CONCLUSION

This study presents a novel speech emotion recognition framework leveraging a hyperparameter-tuned Long Short-Term Memory (LSTM) network that operates directly on raw speech features extracted by MFCC without relying on manually engineered acoustic parameters. The main aim of the research was to create a powerful and efficient model that could enhance the prediction/classification accuracy and overcome the main limitations of the current models, including features redundancy, poor generalization, sub-optimization of the accuracy. The results of the experiment prove that the proposed model can provide significant enhancement in the performance indicators, which proves that it is effective in modeling complex trends in the data. Specifically, the combination of sophisticated feature selection, feature weighting schemes helps a lot to improve the discriminative ability, which is in line with the desired research objectives.

Nevertheless, on closer examination of the findings, some trade-offs can be seen. Although the model is more accurate and higher predictive performance, it is more complicated in terms of computation since it uses deep learning elements. Also, despite the good results of the model on the chosen dataset, the ability to generalize the model to various and real-life datasets is yet to be confirmed. The evaluation is also not as comprehensive due to the lack of such statistical validation and system-level performance measures.

Nevertheless, the conclusions made in spite of these drawbacks show that the suggested method is a valuable improvement to the current techniques as it provides a fair balance between the precision and optimization of features. The findings confirm the central hypothesis that an advanced feature processing technique alongside a multi-level learning framework leads to the improvement in the overall model performance. Future directions need to concentrate on enhancing the scalability, computational overhead, and expanding the assessment criterion, such as real-time applicability, scalability and robustness in dynamical conditions. These enhancements will also increase the

feasibility of the suggested model. The system was evaluated on five publicly available multilingual emotional speech datasets, covering eight distinct emotion classes. Experimental results demonstrate that the proposed LSTM-based model significantly outperforms conventional algorithms in terms of classification accuracy and generalizability across language systems. Experimental results showed that RAVDESS dataset achieved the highest accuracy 98.2% for emotion classification.

Future research will focus on enhancing the real-time capabilities of the proposed model by exploring lightweight LSTM variants and hybrid models such as LSTM-CNN architectures. Model quantization and pruning techniques will be investigated to reduce the computational footprint for efficient deployment on mobile and edge devices. Additionally, transfer learning and domain adaptation methods will be explored to improve robustness on unseen, spontaneous, and noisy speech samples. Moreover, incorporating additional modalities—such as facial cues or physiological signals—may further enrich emotion classification, paving the way for robust emotion-aware AI applications in healthcare, education, and human-computer interaction.

Open Research Issues

Although good results are obtained in this study, there are still many open research issues that can be further investigated. These challenges can be tackled to a great extent to increase the strength, scalability, and applicability of the proposed model to the real world. Generalization across different and large-scale data sets is among the major unresolved questions. Although the proposed model is quite effective on the chosen dataset, its effectiveness on heterogeneous and real-life datasets with different distributions is still to be investigated. The future research should seek to establish the validity of the model in multi-source and cross-domain datasets to make it generalizable.

Computational efficiency and model scalability is another important problem. The combination of deep learning architectures and feature processing mechanisms raises the computation cost, and it is difficult to deploy them to resource-constrained settings. The creation of lightweight models or the use of model compression and optimization methods is an open research field. Noise, incomplete or adversarial data conditions are also yet to be completely discussed in the study. In practice, data is frequently noisy, or can be missing, or can be distorted deliberately. One important research direction is to increase the resilience of the model to such uncertainties. Also, the model is not as interpretable and explainable. Despite the fact that the suggested methodology enhances the effectiveness of predictions, it lacks the adequate information on how decisions are made. Recent research should focus on explainable AI methods in order to render the model more comprehensible and confidence-inspiring, especially when it comes to sensitive areas of application.

The other gap is that of not having comprehensive evaluation metrics. Although conventional performance measurements are applied, another aspect like energy consumption, communication overhead and statistical significance testing has not been included. The inclusion of these metrics would give more comprehensive assessment of model performance.

Author Contributions

This work is contributed by all the authors. The corresponding author developed conceptualization and methodology. The proposed model was implemented and preprocessed by collaborative means of data collection. The respective author made software development and experimental analysis. All the authors participated in validation and interpretation of the results to make sure they were accurate and consistent. The drafting of the manuscript was done by the corresponding author, whereas reviewing and editing was done by all co-author.

REFERENCES

[1] Z. Huijuan, Y. Ning and W. Ruchuan, "Improved Cross-Corpus Speech Emotion Recognition Using Deep Local Domain Adaptation," in Chinese Journal of Electronics, vol. 32, no. 3, pp. 640-646, May 2023, doi: 10.23919/cje.2021.00.196.

[2] L. -M. Zhang, G. W. Ng, Y. -B. Leau and H. Yan, "A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering," in IEEE Access, vol. 11, pp. 71224-71234, 2023, doi: 10.1109/ACCESS.2023.3294274.

[3] B. T. Atmaja and A. Sasou, "Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition," in IEEE Access, vol. 10, pp. 124396-124407, 2022, doi: 10.1109/ACCESS.2022.3225198.

[4] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim and A. Alqahtani, "Mel-MViTv2: Enhanced Speech Emotion Recognition With Mel Spectrogram and

Improved Multiscale Vision Transformers," in IEEE Access, vol. 11, pp. 108571-108579, 2023, doi: 10.1109/ACCESS.2023.3321122.

[5] A. Amjad, S. Khuntia, H. -T. Chang and L. -C. Tai, "Multi-Domain Emotion Recognition Enhancement: A Novel Domain Adaptation Technique for Speech-Emotion Recognition," in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 528-541, 2025, doi: 10.1109/TASLP.2024.3498694.

[6] C. Zhang and L. Xue, "Autoencoder With Emotion Embedding for Speech Emotion Recognition," in IEEE Access, vol. 9, pp. 51231-51241, 2021, doi: 10.1109/ACCESS.2021.3069818.

[7] F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," in IEEE Access, vol. 10, pp. 36018-36027, 2022, doi: 10.1109/ACCESS.2022.3163856.

[8] N. Liu et al., "Transfer Subspace Learning for Unsupervised Cross-Corpus Speech Emotion Recognition," in IEEE Access, vol. 9, pp. 95925-95937, 2021, doi: 10.1109/ACCESS.2021.3094355.

[9] Mustaqeem, M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," in IEEE Access, vol. 8, pp. 79861-79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[10] J. Ancilin and A. Milton, ``Improved speech emotion recognition with Mel frequency magnitude coefficient," Appl. Acoust., vol. 179, Aug. 2021

[11] Lakshman Narayana, V., Rao, G.S., Gopi, A.P., Lakshmi Patibandla, R.S.M. (2022). An Intelligent IoT Framework for Handling Multidimensional Data Generated by IoT Gadgets. In: Al-Turjman, F., Nayyar, A. (eds) Machine Learning for Critical Internet of Medical Things. Springer, Cham. https://doi.org/10.1007/978-3-030-80928-7_9

[12] S. S. Chouhan, A. Kaul, U. P. Singh, and S. Jain, ``Bacterial foraging optimization based radial basis function neural network (BRBFNN) for identification and classification of plant leaf diseases: An automatic approach towards plant pathology," IEEE Access, vol. 6, pp. 88528863, 2018.

[13] M. Lugger and B.Yang, ‘‘The relevance of voice quality features in speaker independent emotion recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Apr. 2007, pp. IV-17–IV-20.

[14] S. Kim et al., “Speech Emotion Recognition Using Graph Attention Networks,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.

[15] J. Lee et al., “Exploring Speaker Information for Speech Emotion Recognition Using Siamese Neural Networks,” in IEEE Transactions on Affective Computing, 2023.

[16] V. Pavani, S. Sri. K, S. Krishna. P and V. L. Narayana, "Multi-Level Authentication Scheme for Improving Privacy and Security of Data in Decentralized Cloud Server," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 391-394, doi: 10.1109/ICOSEC51865.2021.9591698.

[17] J. Park et al., “Temporal Convolutional Networks for Speech Emotion Recognition,” in IEEE Transactions on Affective Computing, 2022.

[18] W. Liang et al., “Attention Mechanism for Speech Emotion Recognition in Noisy Environments,” in Proceedings of IEEE International Conference on Multimedia & Expo (ICME), 2023.

[19] Y. Guo et al., “Speech Emotion Recognition with Multi-level Attention Networks,” in IEEE Transactions on Affective Computing, 2023.

[20] L. Huang et al., “Speech Emotion Recognition with Knowledge Distillation from Pre-trained Models,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.

[21] Q. Yang et al., “Transfer Learning for Speech Emotion Recognition: A Systematic Review,” in IEEE Access, 2023.

[22] Y. Liu et al., “Meta-learning Approaches for Few-shot Speech Emotion Recognition,” in Proceedings of IEEE International Conference on Multimedia & Expo (ICME), 2024.

[23] J. Zhu et al., “Speech Emotion Recognition with Temporal Graph Convolutional Networks,” in IEEE Transactions on Affective Computing, 2022.

[24] Sathiyabhama, B.; Kumar, S.U.; Jayanthi, J.; Sathiya, T.; Ilavarasi, A.K.; Yuvarajan, V.; Gopikrishna, K. A novel feature selection framework based on grey wolf optimizer for mammogram image analysis. Neural Comput. Appl. 2021, 33, 14583–14602.

[25] Dey, A.; Chattopadhyay, S.; Singh, P.K.; Ahmadian, A.; Ferrara, M.; Sarkar, R. A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition. IEEE Access 2020, 8, 200953–200970.